Before we start,

Ki Hyun Kim

nlp.with.deep.learning@gmail.com

Working Process

문제 정의
- 단계를 나누고 simplify
- x와 y를 정의

데이터 수집
- 문제 정의에 따른 수집
- 필요에 따라 레이블링

데이터 전처리 및 분석
- 형태를 가공
- 필요에 따라 EDA 수행

알고리즘 적용
- 가설을 세우고 구현/적용

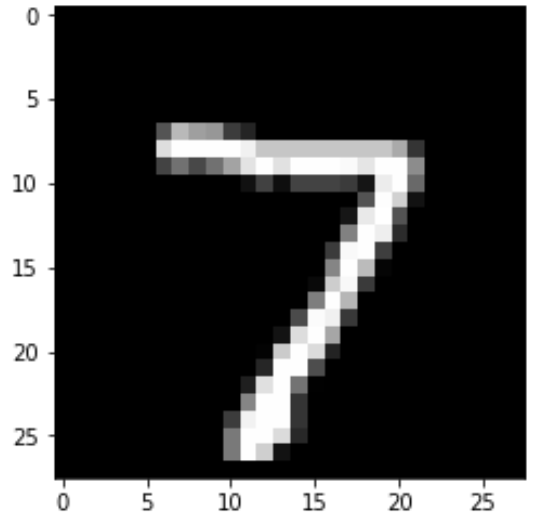
평가
- 실험 설계
- 테스트셋 구성

배포
- RESTful API를 통한 배포
- 상황에 따라 유지/보수

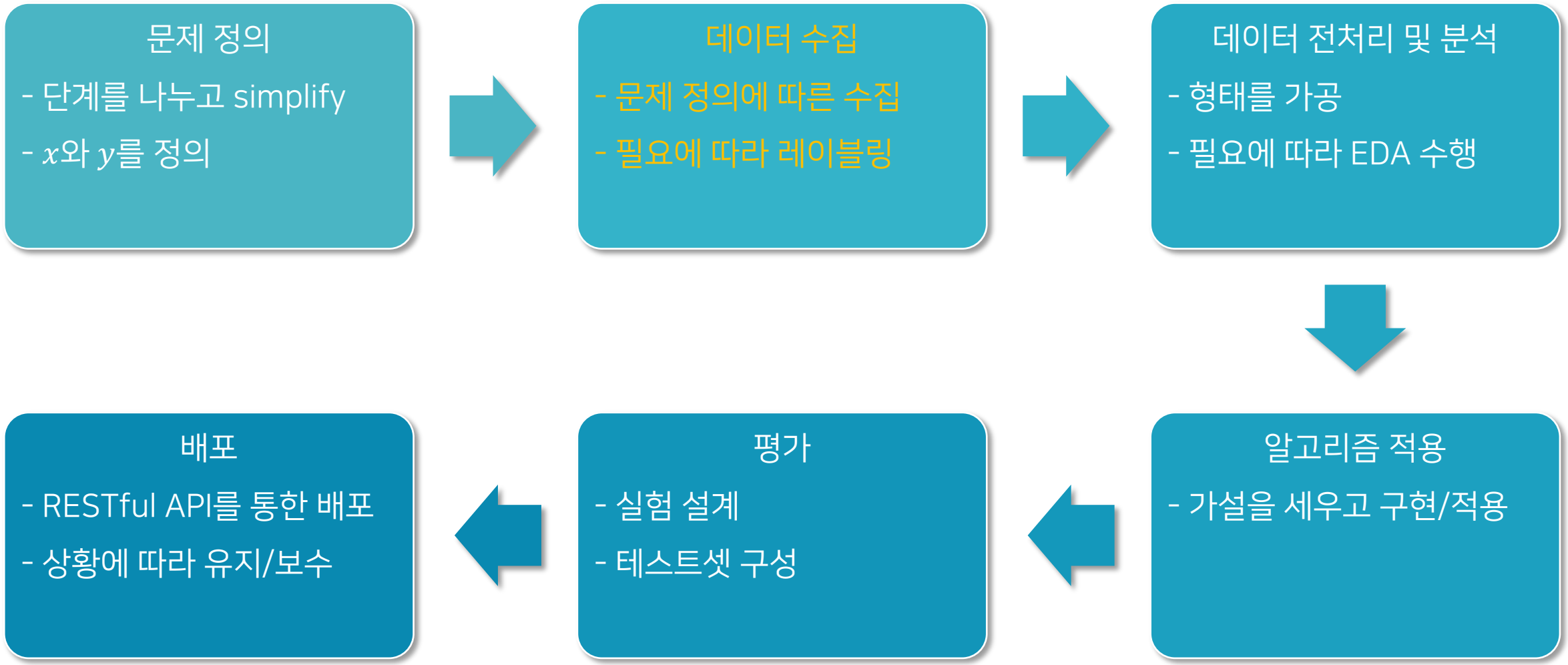
What we are going to do is

- 손 글씨 숫자 한 개를 입력을 받아, 어떤 숫자인지 알고 싶다.
- 글씨와 레이블(label)을 수집하자! → MNIST Dataset

- 입력:
  - 28×28 Grayscale Image
- 출력:
  - 0~9 중의 숫자



Working Process



If you don’t have labels

- 손 글씨 데이터는 모았는데, 레이블(label)이 없다?
  - 직접 labeling 작업을 수행
  - 외주(또는 단기 계약직 고용)를 맡긴다.

- 직접 labeling을 수행할 경우, 대략적인 시간을 계산해보자.
  1) 손 글씨 데이터 70,000장
  2) 파일 이름이 순서대로 적힌 excel 파일을 구성하자. – 효율성이 핵심!
  3) 그림이 순서대로 display 될 때, excel sheet에 label을 적어 넣자.
  4) 기대되는 labeling 속도: 5 secs / 1 sample
  5) 따라서 전체 labeling을 위한 예상 소요 시간: 70,000 * 5 secs = 97.22 hours
  6) 팀원 5명이 고통 분담하자! → 약 20시간 소요 예상
  7) 3일이면 할 수 있다! (3일 후 멘탈 상태는 보장 안됨)

If you don’t have labels

- 손 글씨 데이터는 모았는데, 레이블(label)이 없다?
  - 직접 labeling 작업을 수행
  - 외주(또는 단기 계약직 고용)를 맡긴다.

- 외주를 줄 경우 견적을 내보자.
  1) Task의 난이도에 따라 샘플당 몇 십원에서 많게는 몇 백원 소요
  2) 지금은 쉬우니까: 50원/1샘플 가정
  3) 예상 견적: 70,000장 * 50원 = 3,500,000원
  4) 외주를 주더라도 관리 업무를 위한 자원 필요

- 단기 계약직을 고용한다면?
  - 품질관리를 위한 관리 업무가 상당하게 필요
  - 단기로 업무를 진행할 경우 직원의 능률 하락 가능성

If you don’t have labels

- 손 글씨 데이터는 모았는데, 레이블(label)이 없다?
  - **직접 labeling 작업을 수행**
  - 외주(또는 단기 계약직 고용)를 맡긴다.

초기 데이터를 통해 POC*를 수행한 후,
비용을 확보하여 외주 발주하는 것도 방법.

*POC: Proof of Concept

Working Process

문제 정의
- 단계를 나누고 simplify
- x와 y를 정의

데이터 수집
- 문제 정의에 따른 수집
- 필요에 따라 레이블링

데이터 전처리 및 분석
- 형태를 가공
- 필요에 따라 EDA 수행

알고리즘 적용
- 가설을 세우고 구현/적용

평가
- 실험 설계
- 테스트셋 구성

배포
- RESTful API를 통한 배포
- 상황에 따라 유지/보수

Split into Training / Validation / Test Set

- Random split 수행
  - Train 0.6 : Validation 0.2 : Test 0.2 비율
  - 또는 Test set은 따로 제작하기도 함

Preprocessing

- Tabular Dataset
  - Null value 제거
  - Standard Scale, Min/max Scale 등 적용

- Image
  - 필요에 따라 Augmentation 수행
  - Cropping, Scaling 적용

- Text
  - 특수 기호 제거 등 cleaning, normalization 수행
  - Segmentation 수행
  - 빈도가 적은 단어 제외

Working Process

문제 정의
- 단계를 나누고 simplify
- x와 y를 정의

데이터 수집
- 문제 정의에 따른 수집
- 필요에 따라 레이블링

데이터 전처리 및 분석
- 형태를 가공
- 필요에 따라 EDA 수행

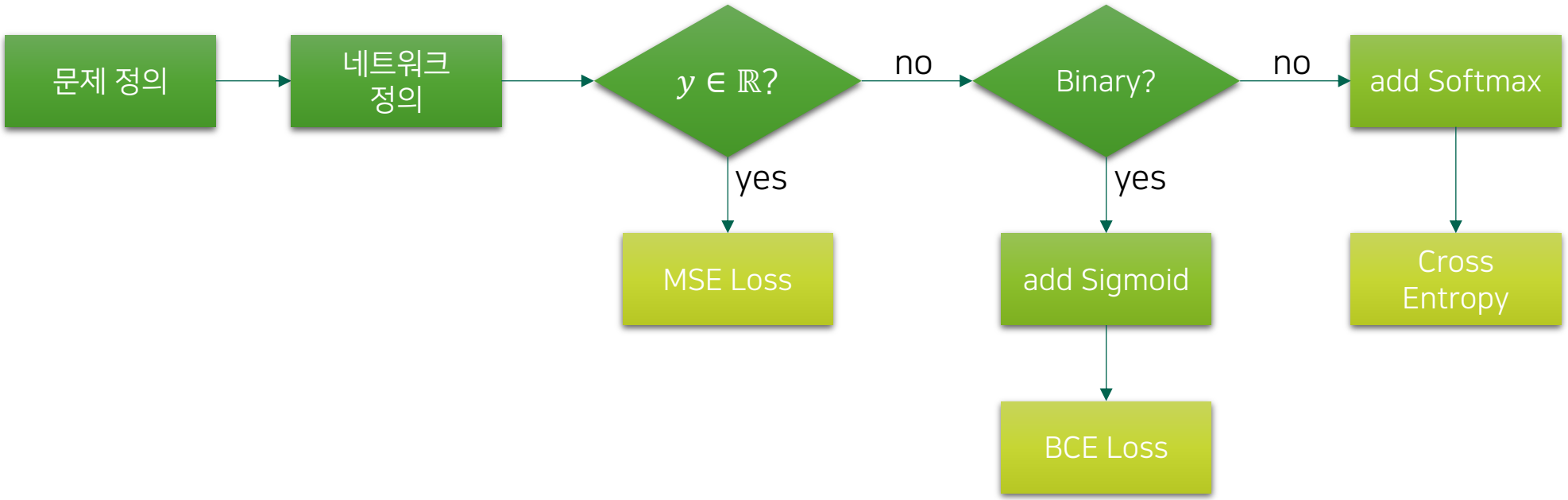
알고리즘 적용
- 가설을 세우고 구현/적용

평가
- 실험 설계
- 테스트셋 구성

배포
- RESTful API를 통한 배포
- 상황에 따라 유지/보수

내부 구성

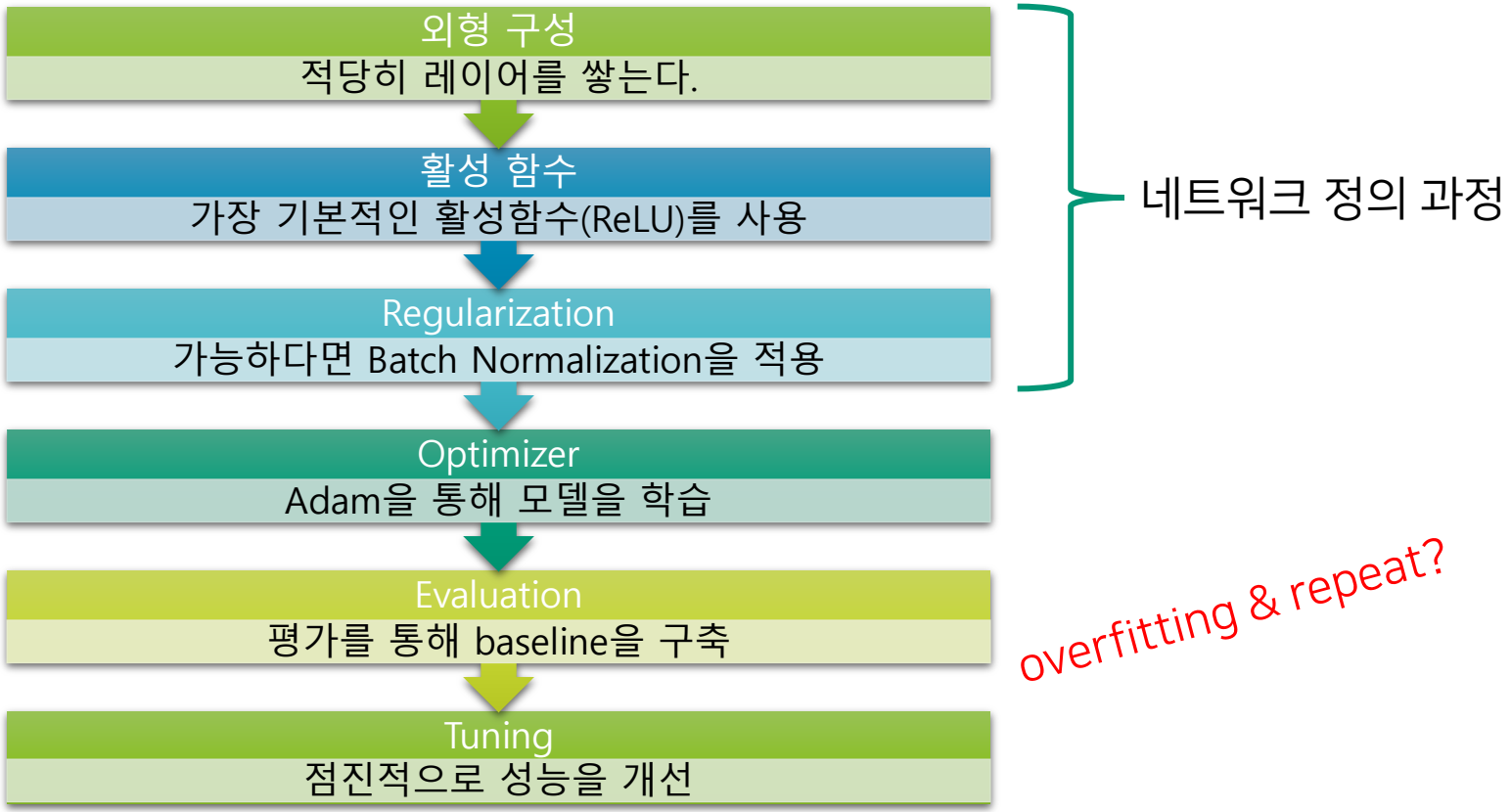
- 문제 정의와 데이터의 성격에 따라 신경망의 구성 요소가 결정



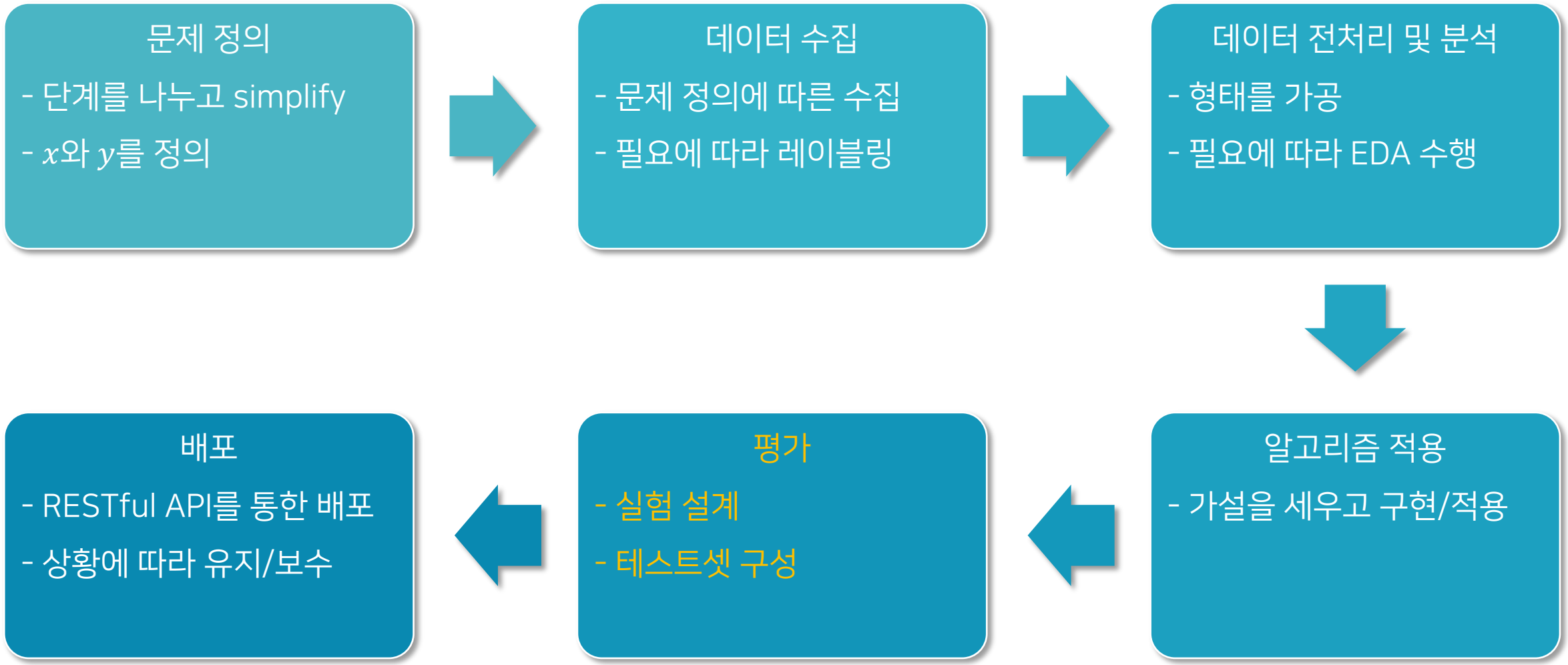
네트워크 설계: ReLU vs LeakyReLU ?

- 레이어의 크기, 깊이 등의 사소한 차이는 큰 성능의 변화를 이끌어내지 않음
  - 풀리지 않던 문제가 갑자기 풀리는 일은 잘 일어나지 않는다!

- 일단 가장 기본적인 형태로 만들어 baseline을 구성하는 것이 중요



Working Process



Hyper-parameter Tuning

- Test set을 대상으로 튜닝을 진행해서는 안됨!

	Train set	Valid set	Test set
Parameter	결정	**검증**	**검증**
Hyper-parameter		결정	**검증**
Algorithm			결정